

XII Simposio Iberoamericano sobre planificación de sistemas de abastecimiento y drenaje

“PREDICCIÓN ON-LINE DE LA DEMANDA DE AGUA URBANA MEDIANTE REGRESIÓN MULTI-KERNEL DE DATOS ESTRATIFICADOS”

Manuel Herrera (1), Joaquín Izquierdo (2), Rafael Pérez-García (3), David Ayala-Cabrera (4)

(1) BATir, Université libre de Bruxelles, Avenue F. Roosevelt, 50 (CP 194/2) B-1050 Bruxelles (Belgique). {mherrera@ulb.ac.be}

(2, 3, 4) Grupo Fluing - Instituto Universitario de Matemática Multidisciplinar (IMM), Universitat Politècnica de València. Camino de Vera S/N, 46022 Valencia, (España). {jizquier, rperez, daaycab}@upv.es

RESUMEN

Este artículo estudia la predicción de la demanda de agua en presencia de una fuente continua de información. Así, el consumo de agua potable puede ser registrado por caudalímetros y enviado por radio-frecuencia a una base de datos central para su almacenamiento y posterior análisis. Esto plantea nuevos retos en el aprovechamiento de la gestión de datos actualizados a tiempo real. Se propone investigar esta nueva perspectiva on-line de los datos mediante métodos de aprendizaje basados en *multi-kernel* (MKr). Los modelos originados mediante este proceso permiten un uso eficiente de datos heterogéneos, extendiendo el concepto de la regresión por Máquinas de Vectores Soporte (SVM), mediante una combinación de tantos regresores como distintos tipos de datos haya disponibles. Un adecuado preproceso de la información basado principalmente en muestreo estratificado, dará la eficiencia computacional necesaria a la predicción con MKr de nuevos datos de la demanda de agua, sin necesidad de recalcular el proceso completo y manteniendo la precisión del modelo.

Palabras claves: Demanda de agua urbana, métodos de aprendizaje on-line, métodos kernel.

ABSTRACT

This paper focus on water demand prediction in the presence of a continuous source of information: water consumption data are registered by flowmeters, and sent by radio-frequency to a central database for storage and posterior analysis. The use of the available information updated in real time poses new challenges to management. Our proposal is to approach this new on-line perspective by the use of multiple kernel regression (MKr) that allows to manage heterogeneous data, extending the simple support vector regression (SVR) to a combination of kernels from as many diverse types of information as kinds of input data. Appropriate pre-processing, mainly based on stratified sampling, offers to MKr enough computational efficiency to predict new water consumptions that avoids recalculating the whole process while maintaining the accuracy of the model.

Key words: Urban water demand, on-line kernel methods, kernel methods.

SOBRE EL AUTOR PRINCIPAL

Autor I: Manuel Herrera es Licenciado en Estadística (Universidad de Valladolid, España) y Doctor en Ingeniería Hidráulica (Universitat Politècnica de València, España). Actualmente es investigador postdoctoral en el Dpto. de Construcción, Arquitectura y Urbanismo de la Universidad libre de Bruselas (Bélgica). Sus áreas de investigación incluyen el diseño de estrategias de optimización multidisciplinar y el desarrollo de modelos estadísticos y de aprendizaje automático aplicados a la Industria y la Ingeniería. Como resultado de su investigación, es co-autor de 16 artículos en revistas internacionales de alto impacto, 5 capítulos de libro con editorial internacional y más de 90 comunicaciones en congresos (actas con ISBN).

INTRODUCCIÓN

El factor de mayor importancia en la planificación y operación de un sistema de abastecimiento de agua es la satisfacción de la demanda del cliente. Ésta conlleva realizar una provisión sin cortes con la adecuada calidad y presión en los distintos puntos de la red. Dentro del marco de la gestión eficiente del sistema de abastecimiento, el cumplimiento de estos objetivos se fundamenta en disponer de predicciones a corto plazo de la demanda de agua. En este sentido, la predicción de la demanda de agua se ha convertido en una herramienta esencial para el diseño, operación y gestión de los sistemas de abastecimiento de agua: es clave en la planificación y desarrollo de futuras ampliaciones de la red, estimando el tamaño y la operación de los depósitos de agua, las estaciones de bombeo y las capacidades de las tuberías (Zhou *et al.*, 2002). Desde otro punto de vista, la predicción de la demanda es necesaria para un adecuado establecimiento de tarifas o para mitigar, en caso de que fuera necesario, los inconvenientes de las interrupciones programadas del suministro. Así, el desarrollo de métodos predictivos es una aplicación de gran auge e importancia, en éste y otros muchos campos de investigación. Añadir a los modelos obtenidos una predicción precisa en tiempo real (modelos on-line) ofrece respuesta a nuevas inquietudes originadas en la gestión de los abastecimientos de agua por el reciente auge de las ciudades inteligentes o *smart cities*: sensores, medidores y sistemas GPS georreferenciados ofrecen una enorme cantidad de datos, a tiempo real, que deben ser aprovechados en la gestión eficiente de los abastecimientos de agua. Además, los modelos on-line establecen una propuesta confiable ante posibles eventos de contaminación o rotura de tuberías, que puedan requerir una solución tan rápida como fuera posible.

Un primer trabajo sobre la predicción de la demanda data de mediados de los años 80 (Maidment *et al.*, 1986). En él se aplica la metodología clásica de Box-Jenkins para estudiar el consumo diario de agua potable como función de la lluvia y de la temperatura. Con el paso del tiempo, los modelos se han ido adaptando a las nuevas tecnologías y a la mayor facilidad para la obtención de datos. De esta forma ha evolucionado el estudio de series temporales, basado sustancialmente en los clásicos modelos ARIMA, a la aplicación de metodologías fundamentadas en Aprendizaje Automático (Shertsa & Solomatine, 2006; Khan & Coulibaly, 2006; Msiza *et al.*, 2007). Herrera *et al.*, 2010, expusieron una interesante comparativa de modelos de

predicción de la demanda basados en diferentes metodologías de Aprendizaje Automático. Como resultado, obtuvieron que la regresión basada en máquinas de vectores soporte (SVMs) obtenía un mejor desempeño en su comparación con otras alternativas como redes neuronales o splines de regresión multivariante, entre otras. En este sentido, los avances en la última década sobre modelos de predicción de la demanda urbana de agua han sido significativos. Sin embargo, la mayoría de los modelos encontrados en la literatura son diseñados off-line, por lo que no representan, de manera adecuada, el estado actual del sistema de abastecimiento de agua para fines operativos, especialmente en casos de emergencia (Machell *et al.*, 2009). Otra desventaja de los modelos off-line reside en su actualización: ya que si su objetivo es una predicción a corto plazo, no podrán mantener su precisión si han de tratar una fuente continua de información (Preis *et al.*, 2009). De esta manera, el modelo deberá recalcularse por completo en periodos cortos de tiempo, lo que conlleva un mayor gasto computacional y una respuesta más lenta en la predicción.

Este artículo propone una solución on-line a un método de predicción de la demanda capaz de trabajar con diferentes tipos de datos, como es la regresión *multi-kernel* (MKr). Esta regresión se entiende como extensión de la basada en SVMs, pero mediante una matriz kernel capaz de recoger información de diferente naturaleza. Esto es una ventaja, ya que se ajusta más a las necesidades usuales en la predicción de la demanda de agua, dado que habitualmente se dispone de datos de naturaleza tan dispar como la fecha, datos climáticos y un histórico de consumo de agua. Los modelos MKr son más precisos aún que los basados en SVMs y hacen un uso más adecuado de los datos. Herrera *et al.*, 2013 propusieron una versión on-line de MKr, basada en la gestión adecuada de las ventanas de tiempo para controlar el error de la predicción sin un aumento significativo del cálculo del modelo. Este trabajo añade a esta actualización de los datos, un paso previo basado en la construcción de diferentes clústeres, donde llevar a cabo tantas regresiones como grupos existan. De esta manera el modelo se adapta de forma precisa a la nueva información disponible, teniendo una respuesta rápida y eficiente ante una información continua.

La siguiente sección introduce la metodología MKr para datos heterogéneos. La propuesta teórica de este trabajo se completa con la propuesta de una metodología on-line de MKr en la siguiente sección. A continuación se presenta un caso-estudio basado

en datos reales de la demanda horaria de agua, del que se discutirán los resultados obtenidos y gracias al cual se expondrá un número de conclusiones y trabajos futuros.

REGRESIÓN MULTI-KERNEL

Los métodos de aprendizaje habituales que se basan en funciones kernel (Schölkopf & Smola, 2001) usan una aplicación implícita del espacio de entrada a un espacio de alta dimensionalidad (denominado espacio de características), definida por una función kernel. Un ejemplo básico de dicha aplicación es el producto interior de dos puntos del espacio de entrada. Se puede hacer uso de otras funciones, y su necesidad viene dada por el objetivo común de convertir relaciones no-lineales complejas en el espacio de entrada en relaciones lineales en el espacio imagen al que conduce la función kernel. Para llevar a cabo el aprendizaje de las características no es necesario una expresión explícita del producto escalar en este espacio final, lo que simplifica los cálculos de los algoritmos asociados con funciones kernel. La Tabla 1 muestra las funciones kernel más comunes.

Tabla 1. Breve lista de las funciones kernel más comunes

Kernel	Expresión
Gaussiano	$K(x, x') = \exp\left(-\frac{\ x - x'\ ^2}{2\sigma^2}\right)$
ANOVA	$K(x, x') = \sum \exp\left(-\sigma(x^k - x'^k)^2\right)^d$
Lineal	$K(x, x') = x^T x' + d$
Polinomial	$K(x, x') = (\alpha x^T x' + d)^d$
Cuadrático racional	$K(x, x') = 1 - \frac{\ x - x'\ ^2}{\ x - x'\ ^2 + d}$

El uso de métodos de aprendizaje basados en funciones kernel trata por sí mismo de integrar datos de varios tipos que son convertidos, de manera automática, a un formato común. Eventualmente, se pueden combinar varias funciones kernel mediante su suma ponderada para después, usando esta función específica, aplicar el esquema clásico de la regresión sobre máquinas de vectores soporte (SVR) (Sonnenburg *et al.*, 2006). Los fundamentos de SVR son resumidos a continuación.

Introducción a la regresión sobre máquinas de vectores soporte

La característica principal de las SVR es que permiten especificar un cierto margen, ε , dentro del cual están dispuestas a aceptar errores. De hecho, el predictor de las SVR se define por esos puntos que caen fuera de la región formada por la banda de anchura 2ε , alrededor de la regresión. Estos son los llamados *vectores soporte* (ver Fig. 1).

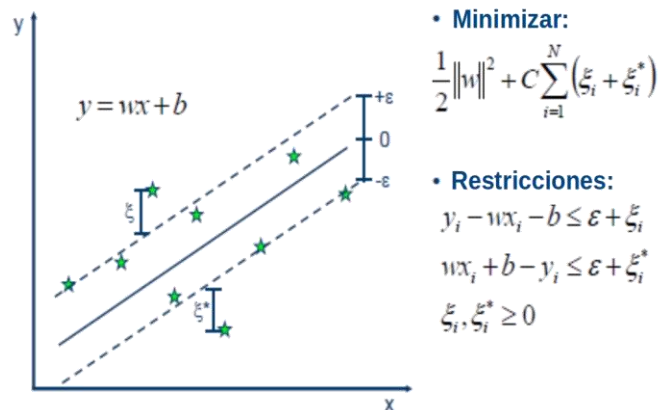


Figura 1. Resumen gráfico y sintético del proceso de obtención de la SVR

El objetivo es encontrar una función $\hat{f}(x)$ que como mucho se desvíe ε del input observado y_i , para la regresión de la Ecuación 1, que busque minimizar la expresión que muestra la Figura 1 (lado derecho).

$$\hat{f}(x) = \langle w, \phi(x) \rangle + b \quad (1)$$

Las restricciones que observamos en la Figura 1 (lado derecho), no han de suponer que siempre exista solución para todos los puntos observados con la precisión dada, por lo que se incluyen las variables de holgura, que permiten ampliar el espacio de búsqueda del óptimo. Además de dar flexibilidad a la regresión, usar esta holgura puede encontrar mejores soluciones, permitiendo la existencia de *outliers* (Schölkopf & Smola, 2002). Sobre la expresión a minimizar de la Figura 1 (lado derecho), se pretende disminuir la complejidad del modelo, siendo C un parámetro de regularización que trata de equilibrar esa complejidad y la precisión que alcanza el modelo final, dado por la expresión de la Ecuación 2.

$$\hat{f}(x) = b + \sum w_i K(x_i, x) \quad (2)$$

Regresión multi-kernel

La regresión basada en SVR usa una sola proyección de los datos y , por tanto, una sola función kernel. Si el conjunto de datos tiene una distribución que es localmente variable, o si son datos, de alguna manera, heterogéneos, usar un kernel sencillo puede no recoger de manera adecuada la distribución de los datos. La fusión de varias funciones kernel puede ayudar a resolver este problema (Christmann & Hable, 2012). Como muestra la Ecuación 3, la fusión de funciones kernel se puede expresar como la suma ponderada de M funciones, cuyo resultado es, de nuevo, otra función kernel (Shawe-Taylor & Cristianini, 2006).

$$\begin{aligned}\tilde{K}(x_i, x_j) &= \langle \phi(x_i), \phi(x_j) \rangle \\ &= \mu_1 \langle \phi(x_1), \phi(x_2) \rangle + \dots + \mu_M \langle \phi(x_M), \phi(x_M) \rangle \\ &= \sum_{s=1}^M \mu_s K_s(x_i, x_j) \quad (3)\end{aligned}$$

Podremos resolver el hiperplano de regresión poniendo el resultado de este multi-kernel en la ecuación relativa a las SVR (Ecuación 4).

$$\hat{f}(x) = b + \sum (\alpha_i^+ - \alpha_i^-) \tilde{K}(x_i, x) \quad (4)$$

EL MODELO DE VENTANAS PARA LA MKr ON-LINE BASADA EN ESTRATOS

Un adecuado preproceso de la información basado, principalmente en muestreo estratificado, junto con la adaptación del modelo a los nuevos datos obtenidos, dará la eficiencia computacional necesaria a la predicción con MKr de nuevos datos de la demanda de agua, sin necesidad de recalcular el proceso completo y manteniendo la precisión del modelo. La Figura 2 nos muestra el proceso general a seguir, que comienza en dicho preproceso en la parte off-line del modelo (con los datos de entrenamiento iniciales) y continúa con una propuesta de ventanas de datos de diferentes tamaños. Estas ventanas buscarán adaptarse a los nuevos datos disponibles, a la vez que la MKr actualiza el cálculo inicial de los parámetros correspondientes tanto al modelo, como los pesos de la suma del mismo. Los parámetros son calibrados mediante la búsqueda heurística de *Grid Search*, en su fase off-line y una rápida adaptación estocástica del mismo en la fase on-line (Arfken, 1985).

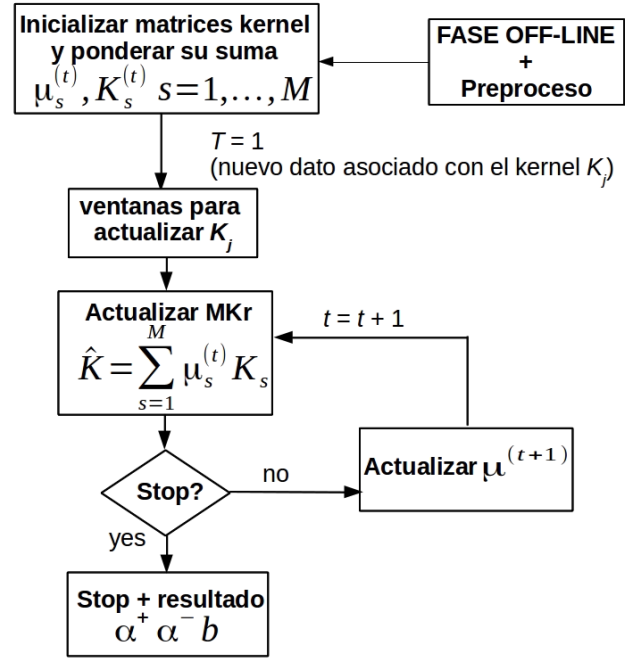


Figura 2. Actualización on-line del modelo MKr mediante una estrategia de ventanas

Preproceso off-line mediante spectral clustering

El trabajo de hacer grupos o clústeres en los datos, previo a su análisis on-line facilita la computación posterior de las diferentes regresiones por estratos, a que disminuye el tamaño muestral. Por otro lado, mantiene o incrementa la precisión con que se ofrecen las predicciones, pues los grupos serán más homogéneos en sus puntos, que los datos tratados globalmente. El análisis de clúster basado en el espectro de una matriz (Ng *et al.*, 2001) es un paradigma de agrupamiento de datos relativamente nuevo. En él se propone usar los autovectores asociados a los autovalores más pequeños (indicativos de partes del grafo fuertemente conectadas) que hayan sido creados por alguna medida de similaridad (Karatzoglou, 2006) o, simplemente, calculados a través de la matriz de afinidad del grafo. De esta forma, planteamos resolver una relajación del problema de partición discreta del grafo, que originariamente es *NP-hard*, y así extender la aplicabilidad del algoritmo clásico de *k-medias*. Este algoritmo tiene la ventaja de poderse adecuar a datos de distinta naturaleza, estando así en sintonía con el resto del trabajo.

Existen dos criterios diferentes para dividir los datos en k clústeres. Uno es usar los dos autovectores asociados a los dos menores autovalores y aplicar sobre

ellos el procedimiento de clustering de manera recursiva, hasta obtener los k grupos. Otro es hacer uso, directamente, de más autovectores de la matriz kernel original. Resumiendo para este último caso: los k primeros autovectores de la matriz de afinidad son usados para formar una matriz $n \times k$, de columnas normalizadas a longitud unidad. Tratando cada fila de esta nueva matriz como un dato usual, podemos usar el algoritmo de k -medias (cualquier algoritmo de clustering será válido en esta fase, tomamos el de las k -medias por simplicidad) para agrupar estos puntos transformados. Las pertenencias de los puntos originales a los clústeres son equivalentes a las pertenencias calculadas sobre los datos transformados.

La Figura 3 muestra el proceso global del algoritmo. Una versión detallada del mismo se encuentra en el trabajo de Ng *et al.*, 2001.

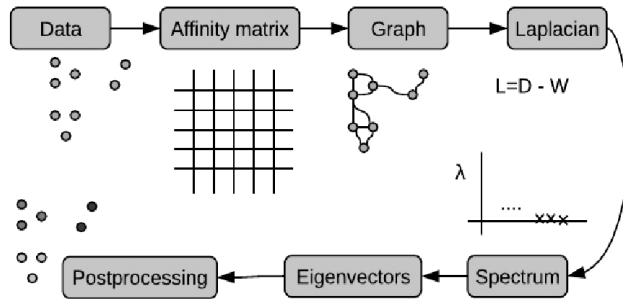


Figura 3. El proceso de spectral clustering

Posterior al *spectral clustering* se realizarán tantas regresiones como grupos se obtengan, para así comenzar la parte on-line de la propuesta mediante la aplicación de una función discriminante a cada nuevo dato que llegue al sistema. De esta manera, cada nuevo punto, x_p , pertenece a un clúster mediante una función basada en las distancias entre ese punto transformado por los kernels correspondientes a los diferentes clústeres $\{\phi_1(\cdot), \dots, \phi_M(\cdot)\}$ y los centroides que definen esos clústeres, $\{a_1, \dots, a_M\}$, como queda especificado en el grupo de Ecuaciones 5.

$$\begin{aligned} & \operatorname{argmin}_m \|\phi_m(x_p) - a_m\| \\ & \operatorname{argmin}_m K_m(x_p, x_p) - 2 \sum \alpha_{j,m} K_m(x_p, x_{j,m}) \\ & + \sum \sum \alpha_{i,m} \alpha_{j,m} K_m(x_{i,m}, x_{j,m}) \end{aligned} \quad (5)$$

Una vez que los nuevos puntos, para los que la regresión se propone, son discriminados en sus respectivos clústeres, se lanzan las estrategias de ventana de actualización on-line de las regresiones que pudieran corresponder (una por cada grupo al

que le han sido asignado nuevos datos). Este algoritmo de movimiento de ventanas para la MKr concreta a dónde se asigne la pertenencia de cada dato, sin que esta operación afecte al resto de la información.

Estrategias on-line para MKr basadas en el movimiento de ventanas de datos

Todas las estrategias de ventanas para realizar una regresión MKr on-line consisten en adaptar el tamaño de la matriz kernel a la llegada de nueva y constante información. De esta manera, los cálculos sólo corresponderán a rehacer la matriz (cálculos de orden n , número de datos) y no a volver a establecer el modelo (cálculo de orden n^3).

Este artículo propone dos alternativas en las estrategias de ventana: *ventanas deslizantes* y *ventanas gusano*, que son una variedad de las anteriores. Las ventanas deslizantes siempre toman la información de los últimos N datos observados. Entonces, cuando observamos un nuevo par (x, y) , contamos con él para la confección de la matriz kernel, pero antes extraemos de esta matriz el par más antiguo (van Vaerenbergh *et al.*, 2006). El grupo de Ecuaciones 6 refleja este proceso.

$$K_j^{(n)} = \begin{bmatrix} K_j^{(n)}(2,2) & K_j^{(n)}(2,3) & \dots & K_j^{(n)}(2,N) \\ K_j^{(n)}(3,2) & & & K_j^{(n)}(3,N) \\ \vdots & & & \vdots \\ K_j^{(n)}(N,2) & K_j^{(n)}(N,3) & \dots & K_j^{(n)}(N,N) \end{bmatrix}$$

$$K_j^{(n+1)} = \begin{bmatrix} K_j^{(n)} & K_j(x_n, x_{n+1}) \\ K_j(x_{n+1}, x_n) & K_j(x_{n+1}, x_{n+1}) + \lambda \end{bmatrix} \quad (6)$$

donde $X_n = (x_{n-N+1}, \dots, x_n)^T$ y λ es un factor de corrección asociado a la regularización de la matriz.

La estrategia de las ventanas gusano consiste en aumentar el tamaño de la matriz kernel cada vez que hay disponibles nuevos datos. La matriz encogerá a su tamaño original cuando su desempeño sea menor que el de una cierta tolerancia. Ese tamaño máximo se calcula entrenando previamente el proceso con datos, que pueden ser ficticios, ya que sólo queremos saber hasta dónde la matriz kernel puede llegar a ser eficiente si sigue creciendo. Este corte cuando la matriz crece mucho, además de ser por motivos computacionales, es adecuado por motivos de sobre-entrenamiento en los datos (Herrera & Filomeno Coelho, 2013). El momento de vuelta al tamaño original también puede ser planificado en

base a los valores de la demanda o a características especiales de la misma (días de fiesta, fines de semana o ciclos de demanda iterativos, por ejemplo, deben contemplarse a la hora de llevar a cabo esta operación). Esta característica de crecimiento de la ventana de datos se ve sensiblemente mejorada, gracias al *clustering* inicial que divide el conjunto de datos inicial y hace partir de tamaños de ventana menores.

La alternativa de las ventanas gusano ofrece una mayor estabilidad en el error cometido por las predicciones del modelo on-line que las ventanas deslizantes, como consecuencia de usar un número de datos mayor o igual que ésta. Sin embargo, las ventanas deslizantes requieren un menor esfuerzo computacional y no hay que entrenar errores asociados al tamaño. Por tanto, dependiendo de los datos y de las necesidades de precisión de los modelos on-line que construyamos, podremos escoger una de estas dos opciones.

CASO DE ESTUDIO

En el presente artículo aplicamos las metodologías anteriores a un sector hidráulico de un municipio del Levante español. Dicho sector cuenta con una extensión de 8 km² y ha de ofrecer un abastecimiento para, aproximadamente, 5000 residentes (consumidores individuales).

La demanda de agua fue medida como la diferencia en las medias entre el caudal entrante y saliente del sector. Esta demanda fue enviada por radiofrecuencia a un centro de base de datos para así proceder a su almacenamiento horario. De esta manera, trabajamos con datos de los consumos de los meses de enero hasta abril del año 2005 (ver Figura 4).

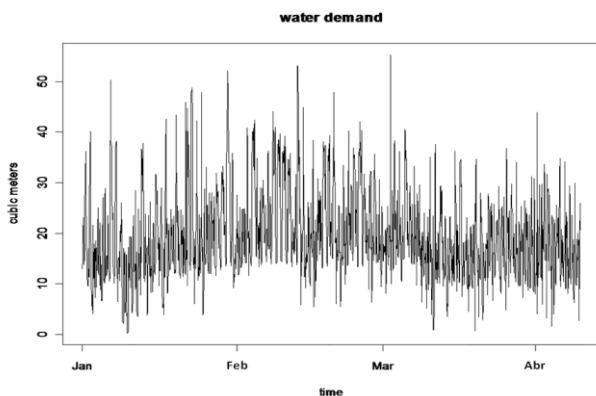


Figura 4. Gráfico de la evolución de la demanda horaria de agua

Además de los valores de consumo de agua, disponemos de información concerniente a datos climáticos diarios: temperatura en Celsius, velocidad del viento en Km/h, milímetros cúbicos de lluvia y presión atmosférica medida en milibares. La Figura 5 muestra cómo se distribuyen estos valores a lo largo de la serie estudiada. Los *boxplot* de esta figura muestran una variabilidad muy reducida de los valores de lluvia, así como una asimetría en las medidas recogidas sobre la velocidad del viento. Los diagramas de series temporales correspondientes revelan la existencia de ciertas tendencias y permiten la observación de una asociación inversa entre los datos de presión atmosférica y temperatura durante estos meses. También es reseñable la baja demanda de agua del inicio de la serie, coincidiendo con bajas temperaturas. Después, ésta crece de manera variable, coincidiendo con el tiempo irregular de los primeros días de la primavera. La influencia de los días lluviosos en el consumo de agua es de especial interés: el día que llueve la demanda promedio aumenta, mientras que baja al día siguiente de la lluvia. Quizá este comportamiento pueda ser atribuido al hecho de que en los días lluviosos una mayor parte de la población se queda en casa durante más tiempo. Todas estas asociaciones, a la vista en la Figura 5, fueron corroboradas aplicando los correspondientes test de correlaciones de Spearman.

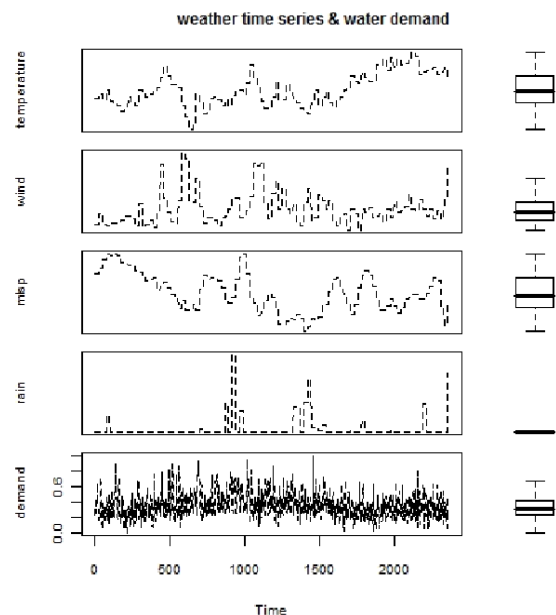


Figura 5. Visualización del impacto de las variables climáticas en la demanda de agua

El objetivo del estudio es aplicar el proceso anteriormente expuesto, sobre esta base de datos y ver su posterior desempeño. Se llevará a cabo una

comparativa de resultados entre las estrategias de ventana con y sin estratificación de los datos, además de buscar el mejor modelo off-line entre las regresiones SVR y MKr. Los datos son divididos como sigue: los 1000 primeros son usados para entrenamiento y validación de los modelos y los siguientes 300 son usados para llevar a cabo la fase test (del modelo off-line, el modelo on-line se explicará después). La combinación kernel utilizada en MKr será: kernel Gaussiano para los datos continuos y kernel Lineal para la parte categórica: día de la semana y hora del día (ver Tabla 1). La comparación es en términos de la raíz del error cuadrático medio (RMSE). De esta manera, para MKr el resultado de RMSE es de 0.10 mientras que para SVR es de 0.13. En ambos casos, los valores de los parámetros fueron calibrados por un algoritmo de búsqueda heurística *Grid Search*. En el caso de MKr son $C = 638$ y $\varepsilon = 0.12$, y la combinación cónica de los kernel Gaussiano y Lineal se realiza mediante los pesos 1.58 y 1.17. Cada uno de estos kernel tiene, a su vez, parámetros $\sigma = 1.37$ para el Gaussiano y $d = 3.55$ para el Lineal (ver Tabla 1). Los valores de MKr predichos (que, como hemos visto, ofrece un menor RMSE que el modelo de SVR), respecto de los 300 nuevos valores obtenidos se representa en la Figura 6.

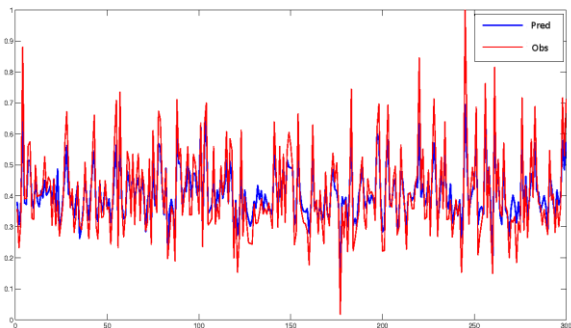


Figura 6. Demanda de agua observada vs. predicha (MKr)

Aunque los resultados del modelo MKr para los siguientes 300 datos son buenos, el modelo va quedando progresivamente obsoleto con la llegada de nuevos datos. El proceso expuesto que hace uso de análisis discriminante más la estrategia de movimiento de ventanas se aplica para los siguientes 1000 datos.

La estratificación de los datos mediante *spectral clustering* obtiene una mejor configuración, respecto del valor promedio de la silueta, con la formación de 3 grupos. Así, habrá un grupo que recoge picos altos de la demanda de agua (formado por 279 elementos de los 1000 de la base de datos de entrenamiento),

otro que hace lo propio con los valores más bajos (116 valores de 1000) y un grupo, más numeroso, formado por los valores intermedios (formado por 605 elementos de los 1000). Respecto de esta configuración de 3 grupos y según simulaciones del esfuerzo computacional de actualizar el modelo con una matriz kernel excesivamente grande, la estrategia on-line de ventanas gusano encogerá (teóricamente) el tamaño de las mismas, a sus medidas originales, cada 500, 200 y 400, nuevos datos según los estratos de clústeres de valores bajos, medios y altos, respectivamente. De esta manera, para los siguientes 1000 nuevos datos, podría resultar, incluso, que ninguno de los tamaños de las ventanas gusano tuviera que ser recortado, lo que indica la ventaja en precisión (se usarán siempre más datos y en regresiones de valores más homogéneos) del uso de una primera parte de *clustering*, previo al proceso on-line. En el caso-estudio este proceso de vuelta al tamaño original de la matriz kernel se ha realizado 2 veces, las dos son en relación a los datos del estrato de valores intermedios de la demanda. En la Figura 7, vemos que esta estrategia permite controlar el error de las predicciones del modelo, sin necesidad de recalcularse nuevamente toda la regresión y con operaciones de orden de complejidad proporcional al tamaño de la ventana de datos usada.

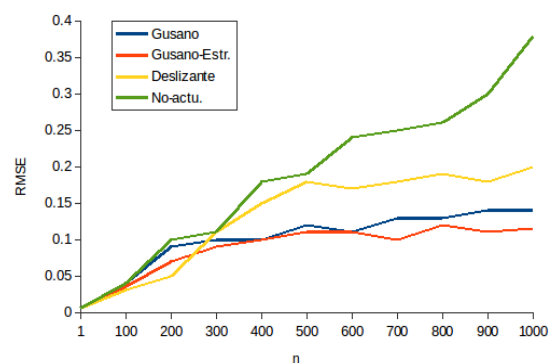


Figura 7. Demanda de agua observada vs. predicha (MKr)

En la Figura 7 se ve cómo crece el RMSE, con los 1000 datos disponibles de test on-line, usando el modelo MKr estratificado sin actualizar. Las estrategias de ventana amortiguan ese crecimiento del error, disminuyendo la necesidad de recalcularse todo el modelo. Incluso el RMSE parece controlarse en el caso del uso de las estrategias de ventanas gusano. En estas últimas se comparan los resultados usando o no la estratificación inicial, obteniendo un mejor resultado si dividimos los datos antes de poner en marcha el aprendizaje on-line.

CONCLUSIONES

El concepto de ciudad inteligente o *smart city* está en continua expansión. De esta manera, son necesarias nuevas perspectivas en la gestión de los recursos de la ciudad dado que la respuesta a tiempo real ante diferentes escenarios que la ciudad afronta se convierte en una obligación en su gestión. Además, esa respuesta debe ser propuesta de manera precisa y haciendo uso de toda la información disponible en nuestros sistemas. Indudablemente, la gestión de una red de abastecimiento de agua y el conocimiento de la demanda de agua urbana siguen esta nueva tendencia, dentro del marco de las *smart cities*. Métodos basados en Aprendizaje Automático, tales como los MKr o SVR introducidos en este trabajo, han emergido estos últimos años como una atractiva opción para las operaciones de predicción y clasificación en abastecimientos de agua debido a su tiempo de ejecución rápido y su facilidad para adaptarse a nuevos escenarios de trabajo.

Como herramientas de aprendizaje on-line, las estrategias de movimientos de ventanas sobre la matriz kernel de MKr, junto con su tratamiento por grupos de datos o clústeres, hace un uso eficiente de toda la información que genera hoy en día una ciudad. La respuesta de esta estrategia es inmediata, dado que no necesita replicar el modelo cada vez que se dispone de nuevos datos. Además, la estratificación de la población, proporciona agilidad y mayor precisión en las predicciones de demanda de agua. Los nuevos retos sobre los que se está trabajando en estas estrategias ahondan en la calibración de los parámetros y en una modificación de las ventanas gusano, permitiendo que ellas mismas evalúen automáticamente cuándo deben hacer los movimientos de expansión/contracción, para llevar un mayor control del error de la predicción.

REFERENCIAS

- Arfken, G. (1985). "The Method of Steepest Descents". *Mathematical Methods for Physicists*, Academic Press, pp. 428-436
- Christmann, A., Hable, R. (2012). "Consistency of support vector machines using additive kernels for additive models". *Computational Statistics & Data Analysis* 56, pp. 854–873.
- Herrera, M., Filomeno Coelho, R. (2013). "Windowing strategies for on-line multiple kernel regression". *International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines*, ROKS-2013, Leuven, Belgium. pp. 105–106.
- Herrera, M., Izquierdo, J., Pérez-García, R., Ayala-Cabrera, D. (2013) "On-line learning of predictive kernel models for urban water demand in a smart city". *12th International Conference on Computing and Control for the Water Industry, CCWI2013*
- Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R. (2010). "Predictive models for forecasting hourly urban water demand". *Journal of Hydrology* 387 (1-2), 121–130.
- Karatzoglou, A., Meyer, D., Hornik, K. (2006). "Support vector machines in R". *Journal of Statistical Software* 15 (9), 1–28.
- Khan, M., Coulibaly, P. (2006). "Application of support vector machine in lake water level prediction". *Hydrological Engineering* 11, pp. 199–205.
- Machell, J., Mounce, S.R., Boxall, J.B. (2009). "Online modelling of water distribution systems: a UK case study". *Drinking Water, Engineering and Science* 2, pp. 279–294.
- Maidment, D.; Miaou, S. (1986) "Daily water use in nine cities". *Water Resources Research* 22 (6), pp. 845-851
- Msiza, I., Nelwamondo, F., Marwala, T. (2007). "Artificial neural networks and support vector machines for water demand time series forecasting". *IEEE International Conference on Systems, Man and Cybernetics*, pp. 638–643.
- Ng, A. Y.; Jordan, M.I.; Weiss, Y. (2001) "On spectral clustering: Analysis and an algorithm", *Advances in Neural Information Processing Systems* 14, pp. 849–856
- Preis, A., Whittle, A., Ostfeld, A. (2009). "Online hydraulic state prediction for water distribution systems". *World Environmental and Water Resources Congress, American Society of Civil Engineers (ASCE)*.
- Schölkopf, B., Smola, A.J. (2001). "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond". MIT Press, Cambridge, MA, USA.
- Schölkopf, B., Smola, A. J., (2002). "Learning with kernels". MIT Press.
- Shawe-Taylor, J., Cristianini, N. (2006). "Kernel Methods for Pattern Analysis". Cambridge University Press.
- Shrestha, D.; Solomatine, D. (2006) "Machines learning approach for estimation of prediction interval for the model output", *Neural Networks* 19 (2), pp. 225-236
- Sonnenburg, S., Ratsch, G., Schäfer, C. (2006). "A general and efficient multiple kernel

- learning algorithm”. Weiss, Y., Schölkopf, B., Platt, J. (Eds.), *Advances in Neural Information Processing Systems 2006*, MIT Press, Cambridge, MA, pp. 1273–1280.
- van Vaerenbergh, S., Vía, J., Santamaría, I. (2006). “A sliding-window kernel RLS algorithm and its application to nonlinear channel identification”. *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2006*, pp. 789–792.
- Zhou, S. L.; McMahon, T. A.; Walton, A.; Lewis, J. (2002) “Forecasting operational demand for an urban water supply zone”, Journal of Hydrology 259, pp. 189-202